Characterizing a Distribution by Distribution Functions

Ian He

Amateur Explorer of $\mathcal{E}$conϕmetric$\$$

July 24, 2023

Table of Contents

Overview I

One crucial feature of the variables we study in Econometrics is that the outcome (or payoff) of interest is stochastic, random, and uncertain. Our goal is to study the **distribution** behind such outcome, instead of the probability that a specific outcome appears.

We have multiple approaches to characterizing the distribution of a random variable. Those approaches, in my opinion, can be divided into two big groups:

- Probability-based (direct) approach;
- Expectation-based (indirect) approach.

Here my focus is the first one.

Overview II

	Discrete Variable	**Continuous Variable**
Probability	PMF	PDF
Cumulative Probability	CDF	CDF

- **PMF**: probability mass function.
- **PDF**: probability density function.
- **CDF**: cumulative probability function.

Probability Mass Function

Discrete Random Variable

Definition 1

A random variable, X, is a **discrete random variable** if there is a discrete set $\mathcal{X} \subseteq \mathbb{R}$ such that $Pr(X \in \mathcal{X}) = 1$. A **discrete set** $\mathcal{X}$ is a set that contains a finite or countably infinite number of elements.

An **indicator function**, $\mathbb{1}\{\cdot\}$, can be used to construct a discrete random variable. For example, for an event A,

$$\mathbb{1}\{a \in A\} = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{otherwise} \end{cases}$$

Definition of PMF

Definition 2

The **probability mass function (PMF)** of a discrete random variable X is a function $f_X : \mathcal{X} \to [0, 1]$ such that $f_X(x) = Pr(X = x)$, the probability that X equals x.

If f_X is the PMF of X, we can write $X \sim f_X$, which is read as "X follows distribution f_X". When it is clear from the context, we can omit the subscript and write only $X \sim f$.

Properties of PMF

The PMF, $p_i \equiv Pr(X = x_i)$, must satisfy two properties:

- $0 \leq p_i \leq 1$;
- $\sum_i p_i = 1$.

These two properties can be easily proved by the three axioms of probability.

Reporting the PMF I

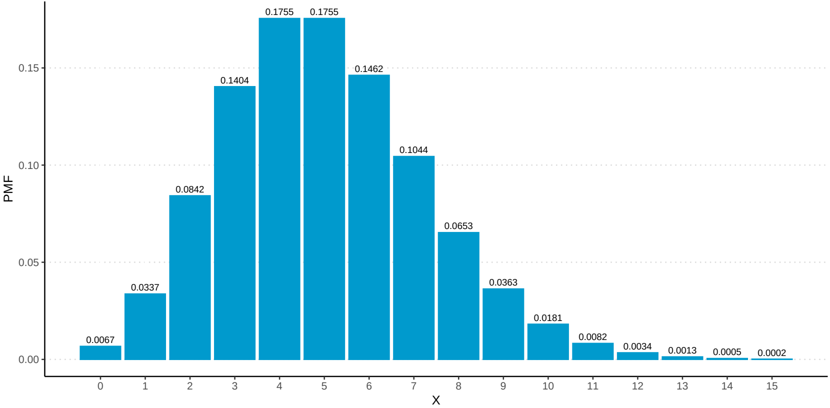
There are three common ways of reporting the PMF (or any probability distribution function) of a random variable:

1) **Mathematical form.** For example, if X follows a Poisson distribution, then its PMF is $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ with parameter $\lambda > 0$.
2) **Table.** For example, a fair coin is flipped until it comes up heads the first time. At that point, the player win 2^n where n is the number of times the coin was flipped. The PMF for possible payoffs is

Payoff	**Probability**
2	$\frac{1}{2}$
2^2	$\frac{1}{2^2}$
$\vdots$	$\vdots$
2^n	$\frac{1}{2^n}$

Reporting the PMF II

3) **Histogram.** For example, the histogram below shows the PMF of a random variable $X \sim \mathit{Poisson}(5)$.



Cumulative Distribution Function

Definition of CDF

Definition 3

The **cumulative distribution function (CDF)** of X is a function $F_X : \mathbb{R} \to [0, 1]$ such that

$$F_X(x) = Pr(X \leq x)$$

If the CDF of X is F_X, we can write $X \sim F_X$, which is read as "X follows the distribution F_X".

Properties of CDF

A CDF has the following unique properties:

- $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$.
- $F_X(x)$ is non-decreasing in x. proof
- $F_X(x)$ is right-continuous: $\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$. continuity
- $F_X(x)$ has a limit from the left, i.e., $\lim_{x\to x_0^-} F_X(x)$ exists.

Note: The above only says that $F_X(x)$ has a limit from the left but doesn't say $\lim_{x\to x_0^-} F_X(x) = F_X(x_0)$. Actually, it is possible that $\lim_{x\to x_0^-} F_X(x) \neq F_X(x_0)$; thus, a CDF may be discontinuous.

CDF versus PMF

For a discrete random variable X, we can obtain the CDF from the PMF by summation:

$$F_X(x) = \sum_{x' \in \mathcal{X} : x' \leq x} f_X(x')$$

Conversely, we can obtain the PMF from the CDF by subtraction:

$$f_X(x) = Pr(X = x) = Pr(X \leq x) - Pr(X < x) = F_X(x) - \lim_{x' \to x^-} F_X(x')$$

Note: CDF is well defined for discrete random variable and non-discrete random variables, but PMF is only defined for discrete random variables (detailed later).

Quantile: The Inverse of CDF

Definition 4

For a $\tau \in (0,1)$, the $100\tau\%$ **quantile** of a random variable $X \sim F_X$, denoted by q_τ or $Q_X(\tau)$, is defined to be

$$F_X^{-1}(\tau) = \inf\{x \in \mathcal{X} : F_X(x) \geq \tau\}$$

where inf is the abbreviation of the "infimum".

As you may know, the inverse function of a CDF F_X doesn't always exist because F_X may not be a bijective (one-to-one) function. However, in the definition of quantile shown above, we use a *specific* inverse function of F_X, which always exists. Actually, there are various algorithms for quantile in statistical software (e.g., there are 9 types in R and type 1 is exactly what we saw above).

Probability Density Function

Continuous Random Variable

Definition 5

A **continuous random variable** is one which takes an infinite number of possible values.

If $X \sim F_X(x)$ and $F_X(x)$ is continuous, then X is a continuous random variable.

PMF Is Not Defined

PMF (defining **absolute probability**) is not defined for continuous variable. Why? Because we always have $Pr(X = x) = 0$ for a continuous random variable X.

Proof:

$$
\begin{aligned}
Pr(X = x) &= \lim_{\epsilon \to 0} Pr(x \leq X \leq x + \epsilon) \\
&= \lim_{\epsilon \to 0} [F_X(x + \epsilon) - F_X(x)] \\
&= 0
\end{aligned}
$$

$\blacksquare$

Thus, we resort to a concept defining **relative probability**: PDF!

Definition of PDF

Definition 6

The **probability density function (PDF)** for a continuous random variable X with CDF F_X is a function that satisfies the following integral equation for all values of $x \in \mathbb{R}$:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt$$

It is clear from the definition of PDF that

$$Pr(a \leq X \leq b) = Pr(X \leq b) - Pr(X \leq a) = \int_a^b f_X(x)\ dx$$

Propeties of PDF

A PDF has the following unique properties:

- $f_X(x) \geq 0$ for all x. This property follows from the fact that $F_X(x)$ is non-decreasing in x.
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$. Geometrically, this means that the area between the PDF and the x-axis must equal 1.

Note: In contract with PMF, PDF $f_X(x)$ can be greater than 1! Recall that a PDF defines the relative probability rather than the absolute probability.

Link between PDF and CDF

Definition 7

A continuous F_X is **absolutely continuous** if it is differentiable at all points in $\mathbb{R}$ except a finite or countably infinite number of them.

If a continuous random variable X has an absolutely continuous CDF $F_X(x)$, then its PDF is of the form

$$f_X(x) = \begin{cases} \frac{dF_X(x)}{dx} & \text{if } x \in \mathcal{X}' \\ \text{arbitrary value} & \text{if } x \notin \mathcal{X}' \end{cases}$$

where $\mathcal{X}'$ is the set of $x \in \mathbb{R}$ at which F_X is differentiable.

Empirical Distribution

Emprical PMF, CDF, PDF, and Quantile I

The definitions of PMF, CDF, PDF, and quantile apply to the **population**. In real life, we cannot observe them, because we only have a **sample** of random variables (i.e., data). Thus, we can only *estimate* them. The estimated ones are called **empirical** PMF, PDF, CDF, and quantile.

Let $(x_1, x_2, ..., x_n)$ be independent and identically distributed samples drawn from some unknown univariate distribution.

- Empirical PMF can be obtained from dividing frequency by number of payoffs:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i = x\}$$

Emprical PMF, CDF, PDF, and Quantile II

- Empirical CDF can be obtained similarly from

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}$$

- Empirical PDF can be obtained by kernel density estimation (KDE):

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

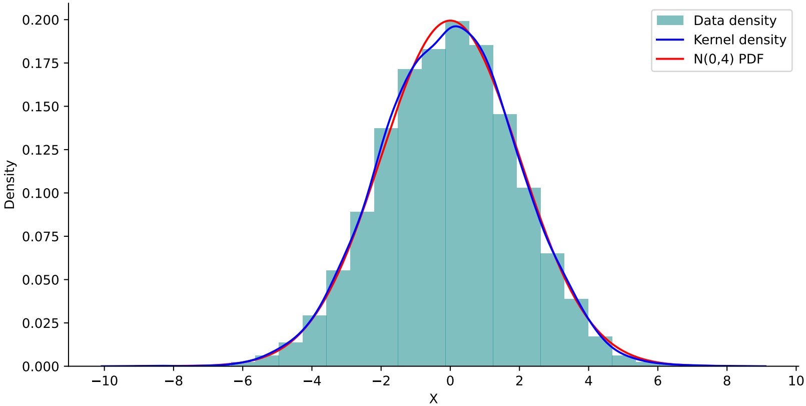
  where h is bandwidth and $K(\cdot)$ is a kernel function.

- Empirical quantile can be obtained from empirical CDF:

$$\widehat{Q}(\tau) = \inf\left\{x : \widehat{F}(x) \geq \tau\right\} = \inf\left\{x : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\} \geq \tau\right\}$$

Emprical PMF, CDF, PDF, and Quantile III

Example: Empirical PDF versus True PDF of $N(0, 4)$



Appendix

Proof of Non-Decreasing CDF

Consider a randon variable X with support set $\mathcal{X}$. Suppose that $x_1 < x_2$ without loss of generality. Set

$$A = \{x \in \mathcal{X} : x \leq x_1\} \quad \text{and} \quad B = \{x \in \mathcal{X} : x \leq x_2\}$$

Then, $A \subseteq B$. Using the implication of probability theory, we have $Pr(A) \leq Pr(B)$; that is,

$$Pr(X \leq x_1) \leq Pr(X \leq x_2)$$

By definition of CDF, we have

$$F_X(x_1) \leq F_X(x_2)$$

Thus, we've proved that CDF is non-decreasing.

back

Limit and Continuity of a Function I

Assume the domain of function f contains an interval (c, d) to the right of c. We say that $f(x)$ has **right-hand limit** L at c, and write

$$\lim_{x \to c^+} f(x) = L$$

if for every number $\varepsilon > 0$ there exists a corresponding number $\delta > 0$ such that

$$|f(x) - L| < \varepsilon$$

whenever $c < x < c + \delta$.

Limit and Continuity of a Function II

Assume the domain of function f contains an interval (b, c) to the left of c. We say that $f(x)$ has **left-hand limit** L at c, and write

$$\lim_{x \to c^-} f(x) = L$$

if for every number $\varepsilon > 0$ there exists a corresponding number $\delta > 0$ such that

$$|f(x) - L| < \varepsilon$$

whenever $c - \delta < x < c$.

Limit and Continuity of a Function III

Let c be a real number that is either an interior point or an endpoint of an interval in the domain of f.

The function f is **right-continuous** at c if

$$\lim_{x \to c^+} f(x) = f(c)$$

The function f is **left-continuous** at c if

$$\lim_{x \to c^-} f(x) = f(c)$$

The function f is **continuous** at c if

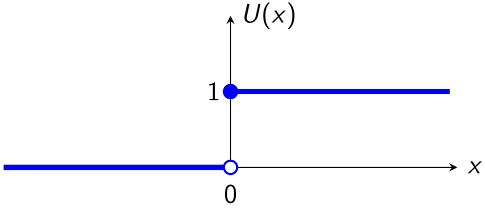
$$\lim_{x \to c^+} f(x) = \lim_{x \to c^-} f(x) = f(c)$$

Limit and Continuity of a Function IV

We say that a function is **continuous** over a closed interval $[a, b]$ if it is right-continuous at a, left-continuous at b, and continuous at all interior points of the interval. This definition applies to the infinite closed intervals $[a, \infty)$ and $(-\infty, b]$ as well, but only one endpoint is involved.

back

For example, $U(x) = \mathbb{1}\{x \geq 0\}$ is right-continuous at $x = 0$, but is neither left-continuous nor continuous there. It has a jump discontinuity at $x = 0$.



References

Hansen, B. E. (2022a).
Econometrics.
Princeton University Press.

Hansen, B. E. (2022b).
Probability and Statistics for Economists.
Princeton University Press.

Hass, J., Heil, C., Thomas, G. B., and Weir, M. D. (2018).
Thomas' Calculus.
Pearson Education, Inc., 14 edition.